Uncertainty in Econometrics: Evaluating Policy Counterfactuals*

Julian Reiss
Centre for Philosophy of Natural and Social Science
London School of Economics

Nancy Cartwright
Department of Philosophy, Logic and Scientific Method
London School of Economics
and
Department of Philosophy
University of California-San Diego

July 2003

*Work on this paper was conducted under the AHRB project *Causality: Metaphysics and Methods*. We are very grateful to the Arts and Humanities Research Board for funding. Nancy Cartwright would also like to thank the Latsis foundation for support. Corresponding address: CPNSS, LSE, Houghton St, London WC2A 2AE, philcent@lse.ac.uk.

1 Introduction

There is without question a great deal of uncertainty in planning and policy formation. Our starting point in approaching this uncertainty is that 'Counterfactuals are the very guide of life'.[1] For rational planning we need counterfactuals because we need to evaluate what would happen were each of the actions under consideration implemented. For this kind of deliberation there is always a vast amount we do not know – facts, laws, probabilities, people's reactions etc. This essay will focus on a more abstract source of ignorance that contributes to uncertainty in planning: We do not know how in principle to evaluate counterfactuals. There is a considerable literature in philosophy on the semantics of counterfactuals and recently contributions in economics itself. The bulk of this literature, we shall argue, puts the problems back to front. It attempts to use counterfactuals to evaluate causal claims; we argue instead that one must use causal claims to assess counterfactuals.[2] That said, there is as yet no general account of how to do so.

The reason the literature has the problems back to front is that parts of both philosophy and economics are still in the grip of the 'Hume problem'. Like David Hume, many philosophers and economists feel that causality is an illegitimate concept. Counterfactuals are called in to secure legitimacy for it. As we shall argue, this skews the study of *genuine* policy counterfactuals – those counterfactuals that can be exploited for policy purposes. The Humean philosophers and economists study a different kind of counterfactuals – counterfactuals that could possibly stand in for causal concepts. But what we need for policy are counterfactuals that describe what would happen were our policies put into place. Unfortunately a semantics geared to the stand-ins for causal concepts fares badly at evaluating policy counterfactuals.

Thus we will argue for two related but independent claims, both concerned with the relationship between causality and counterfactuals. The first is that causal knowledge is required to evaluate policy counterfactuals. The second is that those counterfactuals invoked to stand in for causal concepts are a poor tool for answering the 'What if?' questions policy makers are concerned with.

Section 2 will argue the case that we need causal models to answer the kinds of 'What if?' questions raised in policy and planning. We shall then point out in Section 3 a number of ways in which counterfactuals can be ambiguous. A first step in providing counterfactual hypotheses for use in policy and planning, we maintain, is to be clear exactly what counterfactual hypothesis is under consideration, and this requires explicit assumptions about how the counterfactual antecedent is to be implemented. Section 4 and 5 will describe two attempts to do the job we urge. Section 4 will describe Judea Pearl's theory of causal

[1] Paraphrased from the famous dictum, "Probability is the very guide of life". See Kyburg and Thalos 2003.

[2] Apart from the work on causality and counterfactuals there is also a great deal of work on the logic of counterfactuals. (For a survey, see Edgington 1995.) But the semantics provided in this area is difficult to connect with economic problems. At any rate we shall not discuss it here.

models and counterfactuals, based on his work on Bayes nets. Section 5 will look at the work of James Heckman.

Both of the accounts we shall look at use counterfactuals to explicate causal notions. Neither, however, is engaged in a fully-fledged Hume programme. For both begin, as we urge one must, with a causal model, a model that contains some set of causal principles. The notion of causality involved in the causal principles is left unanalysed. The causal model is used to evaluate counterfactuals and then the causal concept of concern is defined in terms of the counterfactual. Even if this method does not provide a total elimination of causal concepts, if successful it can at least explicate more problematic causal notions in terms of less problematic ones. The cost, however, is in the hope to use the accounts to answer more general kinds of counterfactual questions. The attempt to capture the intended causal notions skews the accounts; whether or not they succeed at explicating the causal notions of concern, neither can double as a semantics for the 'What if?' counterfactuals required for policy formation.

2 The Need for Causal Models to Evaluate Policy Counterfactuals

Policy analysts are often interested in answers to questions with the structure 'What would have happened to Y (the 'target variable'), had X been x (the 'control variable' and its value, respectively)?' Both control and target variable can be binomial or continuous. We might for example ask what would have happened to the profits of British utilities had they not been privatised or whether the 9/11 incidents would have occurred had investments in national security been as high as demanded by some Democrats. We want to call these questions *genuine* 'What if?' questions (and the corresponding policy counterfactuals, *genuine* policy counterfactuals) because although formally they refer to hypothetical scenarios, their answers prepare the policy maker for a situation in which what he supposes actually happens.

How does one find answers to a genuine 'What if?' question? We contend that such questions are always answered against a causal model in which the control variable figures as a cause and the target variable as an effect.

Philosophers, and some economists, tend to regard priority the other way around. Many thinkers in the Humean tradition view counterfactuals as more fundamental and causal relations as derivative. David Lewis, for example, develops a counterfactual theory of causation that takes the following quote from Hume as starting point:[3]

> *we may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second.* Or, in other words *where, if the first object had not been, the second never had existed.*

[3] quoted from Lewis 1993/1973, p. 193

Based on this underlying idea, Lewis develops a theory that defines causal relations in terms of 'chains of counterfactual dependence'. The details of his theory do not matter here. But in our view, it is peculiar to invoke Hume for this project. Hume was an empiricist of a particular brand. For him, a concept was meaningless unless associated with an idea which itself was a copy of a direct sense impression. Now, it is disputable whether we can have direct sense impressions of causal relations. Hume certainly believed that we cannot. But it is absurd to suppose that there is a sense impression from which we can copy the idea of the 'object that would never have existed'.

Hume's associationist theory of knowledge has long been out of fashion. In its stead, a contemporary empiricist will demand that our claims are made on the basis of the best evidence at hand. But from the evidential point of view, counterfactuals do not seem to fare any better. No (direct) evidence can be had for a statement that describes a state of affairs which is 'counter to the facts'. The usual strategy, then, is to translate the counterfactual into a different kind of statement, for example a statement about possible worlds or about laws of nature. Although theories of counterfactuals tend to focus on semantic rather than epistemological issues, eventually the problem of how to justify belief in counterfactuals must be addressed. Robert Stalnaker, for example, notes:[4]

> For similar reasons, the empiricist may be uncomfortable about a theory which treats counterfactuals as literal statements about non-actual situations. Counterfactuals are often contingent, and contingent statements must be supported by evidence. But evidence can be gathered, by us at least, only in this universe. [...]
>
> It is because counterfactuals are generally about possible worlds which are very much like the actual one, and defined in terms of it, that evidence is so often relevant to their truth. When I wonder, for example, what would have happened if I had asked my boss for a raise yesterday, I am wondering about a possible world that I have already roughly picked out. It has the same history, up to yesterday, as the actual world, the same boss with the same dispositions and habits.

But now it seems that the theorist who regards counterfactuals as more fundamental than causal relations is in a dilemma. The statement into which the counterfactual is translated either does or does not involve causal concepts. If it does involve causal concepts, the theory is circular and counterfactuals have not been shown to be more fundamental than causal relations (note for example Stalnaker's own use of the notions 'dispositions' and 'habits'—both are, in our view, causal notions).[5] If, on the other hand, the statement does not involve

[4] Stalnaker 1968

[5] It is a different question whether the circularity is vicious or virtuous. Below, we will discuss a case where certain causal concepts are defined in terms of counterfactuals, which in turn are evaluated in a causal model. But knowledge about the causal concepts defined in this procedure is not required in the model construction. Thus new causal knowledge is extracted

causal concepts, the theory is likely to fall prey to one or more of the difficulties that agonise all reductive theories of causation. These difficulties include[6]:

- the problem of concomitant effects
- the problem of co-extensionality
- Simpson's paradox
- ...

Of course, one cannot prove that all reductive theories, even future ones, will suffer difficulties like these. However, from the past record of failed attempts it seems a reasonable move to give up trying rather than keeping to fail.

Witness that our argument here against counterfactual theories of causation differs from the usual strategy employed when these theories are criticised. The usual strategy consists in choosing a particular formulation of a theory, finding a case where one would intuitively say that x causes y but where the theory yields a negative answer (or the other way around). Bill and Suzy throw rocks at a bottle[7]. Bill's rock hits the bottle a split second before Suzy's does and the bottle shatters. Intuitively, we would say that Bill's throwing the rock caused the bottle to shatter. But had he not thrown the rock, the bottle would have shattered anyway. Thus, according to one naïve formulation of the counterfactual theory, Bill's throwing does not come out as the cause of the shattering. And yet, we would say it is the cause. Thus this version of the theory must be false.

These criticisms, then, are based on intuitions or ordinary language usage. The argument put forward here, by contrast, is based on evidential considerations. It maintains that the evidence in favour of a counterfactual can never be better than the evidence in favour of the associated statement about causal relations. Therefore, we reject the view that counterfactuals are more fundamental.

Thus far, conditional on the soundness of our argument, we can reject counterfactual theories of causation, but this does not give us a causal theory of counterfactuals. Why, then, believe that counterfactuals can only be analysed in terms of causal models? The reasons for this are mainly pragmatic. There is a great variety of kinds of counterfactuals, and each kind demands its own mode of translation. Just consider a few stock philosophical examples:

- If Bizet and Verdi had been compatriots, Bizet would have been Italian.
- If God had had a daughter, it would have been Carol.
- If the pressure in this container had been p, its temperature would have been t.

(among other things) from old causal knowledge. This kind of circularity is not vicious but it renders the counterfactual definition superfluous.

[6]For further discussion see Cartwright 1983 and 1989.

[7]This is case that widely discussed by philosophers. See for example Lewis 2000.

- If Nixon had pushed the button, there would have been a nuclear holocaust.

To answer whether the first counterfactual is true is a matter of logic and meaning. To be compatriots, Bizet and Verdi would have needed to live in the same country. This could have been Italy, France or any other country for that matter. Thus the truth value is indeterminate. But we do not need *causal* knowledge in order to evaluate it. Living in a country does not cause one to be someone else's compatriot. The second counterfactual is metaphorical. In order to answer it, we may want to ask whether Carol has a number of characteristics we conventionally ascribe to Jesus, say, kind-heartedness and generosity. For the third one, we might invoke a model but (according to many) a purely associational model is sufficient. We do not need to know how the pressure value comes about or how it brings about the temperature. Only the fourth counterfactual seems to require causal knowledge, knowledge about the technological and sociological nexus Nixon lived in.

The point is that different kinds of counterfactuals require different kinds of translations. *A priori*, there is no primacy of translation into causal models. But the topic here is not counterfactuals *simpliciter* but counterfactuals relevant to policy making. It seems thus safe to suppose that counterfactuals which can be addressed on the basis of logical or semantic relations, or which are metaphorical in nature do not play an important role.

The matter is not so simple with respect to causal versus associational models. Laws of association (or the models that represent them) *can* be used to evaluate counterfactual claims. Just recall that the gas law – an associational law – was used above with reference to the third example. But in the context of policy making in most cases the formulation of the question is at least partly causal in nature. Usually, we ask whether one can use a certain socio-economic variable as a kind of handle to control another variable – the money stock to control prices, investment in schooling to control education and the latter to influence income and inequality, interest rates to control economic activity. Causal models, then, are the appropriate tools to evaluate these kinds of policy counterfactuals.

There is also another reason. Suppose our counterfactual question was of a kind that could be answered on the basis of an associational law. For instance, we want to know what the unemployment rate would have been had inflation been x – without implying either that we would have *caused* the inflation rate to assume that value or that there is a causal relationship between inflation and unemployment. We may then want to use the Phillips curve to answer that counterfactual. The problem with that suggestion is that the Phillips curve and many other economic 'laws' are at best more or less stable empirical relationships but not ones that deserve the name 'law'. We may calculate measures of association between variables but these do not usually sustain counterfactuals about what would happen under interventions. That such 'laws' fare badly at making policy predictions is a sign of that.

On the other hand, all methods in economics designed to answer counterfac-

tual questions are in fact methods of *causal* (rather than associational) inference. Bayes'-nets methods[8], the potential outcome framework[9], Heckman's 'counterfactuals'[10], natural experiments à la Herbert Simon[11], James Hamilton[12] and others, Kevin Hoover's invariance account[13], they all answer causal questions. Importantly, if they fail to answer these questions (because the method has not been applied correctly or it is inadequate for the situation at hand for instance), they do not answer associational questions either – they answer nothing.

There are no established methods of associational inference in econometrics. So we cannot show that no associational models will work. It is not entirely established, then, that policy counterfactuals require causal models for their proper evaluation but currently this appears to be the best one can do.

What then is a causal model? That, we contend, is the central question. What characterizes a causal model depends on what one wants to do with the model. Both Pearl and Heckman provide explicit answers to the question of what constitutes a causal model and both give explicit rules for how to assess the truth values of counterfactuals given a causal model. Both use the counterfactuals they evaluate to define new causal notions. But neither is adequate in general for assessing genuine counterfactuals of direct use in planning and policy formation. That is why we urge that more work needs to be done.

An adequate account of how to evaluate counterfactuals from causal models will require

- A description of what a causal model consists in
- Rules for how to evaluate counterfactuals given the model
- An argument to show that the results are correct: the propositions thus evaluated really are the ones we are trying to assess.

Pearl provides all three ingredients, as a good formal account of counterfactuals and causal models should. Despite this, we shall argue, the counterfactuals he assesses are not in general the ones we need for policy and planning, though they may serve a variety of other purposes, including the definition of new causal notions. Heckman has an explicit proposal for the first two. As to the third, we shall show that, trivially, his characterization captures the concept of *causal contribution* when applied in linear systems, but it will not double as a semantics for the kinds of counterfactuals that can be put directly to use in planning and policy formation.

[8] See *e.g.* Spirtes *et al.* 2001 and Pearl 2000.

[9] Rubin 1977. For an accessible account, see Holland 1988.

[10] Heckman 2000 and 2001. It will become apparent in a later Section of this paper why we put the counterfactual in scare quotes here.

[11] Simon 1953 and 1954

[12] Hamilton 1994

[13] Hoover 2001. For a discussion, see Reiss forthcoming.

3 Exactly What Counterfactual Question Is at Stake?

The question of 'implementing the antecedent' is hugely important for policy purposes. In the philosophy literature the topic is usually discussed under the heading of 'backtracking'. The dominant view is that counterfactuals have to be non-backtracking. That is, for the evaluation of a counterfactual it should not matter how the antecedent is brought about. However, in general the evaluation of the kinds of genuine counterfactuals of direct use for policy will be extremely sensitive to the methods of implementation. Consider a standard example from philosophy. Jack and Jim had a quarrel yesterday and Jack is still furious. The question is if had Jim asked Jack for a favour today, would Jack have obliged? One plausible answer is 'Yes', since for Jim to ask Jack for a favour, there would have had to have been no quarrel before.[14]

Implementation matters all the more for policy. Suppose the counterfactual question of interest is 'Had investment in schooling, I, been i (rather than $i^* < i$, the actual value), what would the income, Z, of the cohort have been?' In the real world, the additional money cannot be manna from heaven – it has to be raised somewhere. And it might matter whether it would have been taken from the defence budget or the social security budget. For in either case, it is very likely that the cohort would have been influenced by the loss (*e.g.* because some unemployed would have received less benefit or some recruits would not have been hired) but in an asymmetric way (it is plausible to assume, for instance, that the cohort would have been less influenced had the money come from the defence budget).

So to assess the truth value of any particular counterfactual that we hope to use in policy formation we will need to know what changes are supposed to happen, where often the exact details matter. Sometimes when we consider a policy we have a very definite idea in mind how it will be implemented. We shall call the related counterfactuals, 'implementation specific'. At the other end of the scale, we might have no idea at all; the counterfactuals are 'implementation neutral'. When we evaluate counterfactuals, we had better be clear what exactly we are presuming.

For counterfactuals that are totally implementation specific, we know exactly what we are asking when we ask 'What would happen if...?'[15] For others there are a variety of different strategies we might adopt. For one, we can employ the usual devices for dealing with epistemic uncertainty. We might, for instance, assess the probabilities of the various possible methods of implementation and weight the probability of the counterfactual consequent accordingly. In the methodology of economics literature we find another alternative: Stephen LeRoy and Daniel Hausman focus on counterfactuals that would be true *regardless* of

[14] See *e.g.* Lewis 1979.

[15] Or rather, we know this relative to the factors included in the causal model. Presumably no causal model will be complete, so this remains as a source of ambiguity in our counterfactual claims.

how they are implemented. We begin with LeRoy.

LeRoy's stated concern is with causal ordering among quantities, not with counterfactuals. But, it seems, he equates 'p causes q' with 'if p were to change, q would change as well' – so long as we give the 'right' reading to the counterfactual. It is his proposed reading for the counterfactual that matters here. It may help to present his brief discussion of a stock philosophical example before looking to more economic cases – the case of birth control pills and thrombosis.

Birth control pills cause thrombosis; they also prevent pregnancy, which is itself a cause of thrombosis. LeRoy assumes that whether a woman becomes pregnant depends on both her sexual activity and whether she takes pills. Now consider: 'What would happen vis-à-vis thrombosis were a particular woman to become pregnant?' That, LeRoy points out, is ambiguous – it depends on whether the change in pregnancy comes about because of a change in pill-taking or because of a change in sexual activity.

In his formal characterisation LeRoy treats systems of linear deterministic 'reduced form equations': 'In current usage an economic model is a map from a space of exogenous variables – agents' characteristics and resource endowments, for example – to a space of endogenous variables – prices and allocations.'[16] LeRoy assumes that the equations are functionally correct and that variables designated as 'exogenous' are not caused by any of the remaining (endogenous) variables. Since they are functionally related to the endogenous variables, we may assume that either they are causes of some of the endogenous variables or are correlated with such causes. For LeRoy's purposes it seems we must suppose they are causes. He also supposes that the possible sources of implementation for a change in an endogenous variable are exactly the members of the minimal set of exogenous variables that will fix the value of the endogenous variable according to the economic model. Together these assumptions constitute a characterization of the causal model that will be used to evaluate counterfactuals.

For illustration of his semantics, LeRoy considers a familiar supply and demand model:

$$\begin{aligned} q_s &= \alpha_s + \alpha_{sp}p + \alpha_{sw}w \\ q_d &= \alpha_d + \alpha_{dp}p + \alpha_{di}i \\ q_s &= q_d = q. \end{aligned} \tag{1}$$

Here p is price; q, quantity; w, weather; i, income. LeRoy asks what the effect of a change in price would be on the equilibrium quantity. By the conventions just described, a change in price can come about through changes in weather, income or both, and nothing else. But, LeRoy notes, "any of an infinite number of pairs of shifts in the exogenous variables 'weather' and 'income' could have caused the assumed changes in price, and these map onto different values of

[16] LeRoy 2003, p. 1

q".[17] Thus the question has no definite answer – it all depends on how the change in p is brought about.

LeRoy contrasts this model with a different one:

$$\begin{aligned} q_s &= \alpha_s + \alpha_{sw} w + \alpha_{sf} f \\ q_d &= \alpha_d + \alpha_{dp} p + \alpha_{di} i \\ q_s &= q_d = q, \end{aligned} \tag{2}$$

where f is fertilizer. Here fertilizer and weather can change the equilibrium quantity, and no matter how they do so, the change in price will be the same. In this case LeRoy is content that the counterfactual, "If q were to change from taking the value Q to taking the value $Q + \Delta$, p would change from $P = (Q - \alpha_d - \alpha_{di} I)/\alpha_{dp}$ to $P = (Q + \Delta - \alpha_d - \alpha_{di} I)/\alpha_{dp}$" is unambiguous (and true). The lesson he draws is the following (where we substitute counterfactual language for his causal language): "[Counterfactual] statements involving endogenous variables as [antecedents] are ambiguous except when all the interventions consistent with a given change in the [antecedent] map onto the same change in the [consequent]".[18]

The statement as it stands is too strong. Some counterfactuals are, after all, either implicitly or explicitly implementation specific. In (1) we could ask, for instance, what the value of q would have been had $p = P$ been brought about by $i = I$. What LeRoy offers instead is a semantics for counterfactuals that are, either implicitly or explicitly, implementation neutral. In this case the consequent should obtain *no matter what possible change occurs to bring the antecedent about.*

Daniel Hausman seems to have distinguished between implementation-specific and implementation-neutral counterfactuals, too, as we do here, though he does not explicitly say so. He considers an example in which engineers designing a nuclear power plant ask, "What would happen if the steam pipe were to burst?"[19] The answer, he argues, depends on how it will burst. "Responsible engineers", he argues, must look to the origins of the burst "when the consequences of the pipe's bursting depend on what caused it to burst."[20] That is, in these situations responsible engineers will ask implementation-specific 'What if?' questions.

On the other hand, when Hausman turns to providing some constraints that a general semantics for counterfactuals must satisfy if we are to use counterfactuals to establish causal order, he seems to be concerned with implementation-neutral counterfactuals. It is not worth going into the details here; the results are similar to LeRoy's. Any semantics that satisfies Hausman's constraints should give the same result as LeRoy's prescription when restricted to counterfactuals evaluated via what LeRoy calls an 'economic model'.

[17] *ibid.*, p. 6
[18] *ibid.*, p. 6
[19] Hausman 1998, p. 122
[20] *ibid.*

So we may have counterfactuals that are implementation specific; we may have ones that assume some one or another of a range of possible implementations; and we may have implementation-neutral ones where we wish to find out what would happen no matter how the change in the antecedent is brought about. For thinking about policy we had better know which kind of counterfactual we are asserting and ensure that our semantics is appropriate to it.

4 How We Evaluate Counterfactuals: Pearl and Why That's Not Good Enough

We begin with Judea Pearl because his account is the most formal, the most complete and the most powerful of any methods currently available, allowing evaluations of complex counterfactuals and their probabilities. Pearl is famous for his work on artificial intelligence, causal search algorithms and reasoning under uncertainty. In his work on causality, he is an advocate of the so-called Bayes'-nets methods. Although Bayes' nets can be put to a large variety of uses, one prominent application lies in causal inference. Here, they function as a tool for learning causal relationships from conditional probabilities on the basis of assumptions about the system considered.

As tools for causal inference, Bayes' nets methods are part of what we have called the 'Hume programme': the programme that aims at replacing 'problematic' causal notions with 'unproblematic' ones. We must emphasise that Pearl himself is not a strict Humean. He does not want to eliminate causal talk altogether. But he too uses his system in order to explicate more problematic causal notions, for example those relating to singular causation, in terms of less problematic notions, for example that of a causal model. Pearl's semantics for counterfactuals plays a vital role in his explication of causal notions. Since one might be tempted to think it doubles as a semantics for genuine 'What if?' questions, let us examine his account in more detail here.

Pearl envisages counterfactual statements of the form 'The value that Y would have obtained, had X been x'.[21] According to him, counterfactual statements are always evaluated within a causal model. A *probabilistic causal model* is defined as a quadruple $< U, V, F, P(u) >$, where:

1. U is a set of variables that are determined by factors outside the model;
2. V is a set of variables that are determined by variables in the model;
3. F is a set of functions $\{f_1, f_2, \ldots, f_n\}$ such that each f_i is a mapping from $U \bigcup (V \setminus V_i)$ to V_i and such that the entire set F forms a mapping from U to V;
4. $P(u)$ is a probability function over the domain of U.[22]

[21] Pearl 2000, p. 204 (Def 7.1.5 'Counterfactual')

[22] This is actually a merger of two of Pearl's definitions. For the precise formulations, see Pearl 2000, pp. 204f. (Definitions 7.1.1, 'Causal Model', and 7.1.6, 'Probabilistic Causal Model').

Condition (iii) says that each f_i gives use the value of V_i given the values of all other variables in $U \bigcup V$, which is unique in case the system is recursive or 'acyclic'.[23] Each f_i is supposed to be functionally correct in the situation modelled and the quantities on the right-hand-side are supposed to be a complete set of 'direct causes' of the quantity on the left.[24]

An important notion in Pearl's system is that of a *submodel*. A submodel M_x of a model M is relative to realisations x of a set of variables X. It is formed by deleting all functions in M that have members of X as an effect and replacing it with the constant function $X = x$. The effect of action 'set X to x', in short $do(X = x)$, on M is given by the submodel M_x. Finally, a counterfactual of the form 'The value Y that would have obtained, had X been x' on M is defined as the solution of M for Y under the action $do(X = x)$, that is, of M_x. The counterfactual value Pearl also calls 'potential response' and abbreviates it with $Y_x(u)$.

Pearl introduces a theorem according to which a counterfactual can be evaluated using three steps.[25] The theorem shows how to assess the conditional probability $P(B_A|e)$ of a counterfactual statement of the form 'If it were A then B', given evidence e:

1. **Abduction** – Update $P(u)$ by the evidence e to obtain $P(u \mid e)$.
2. **Action** – Modify M by the action $do(A)$, where A is the antecedent of the counterfactual to obtain the submodel M_A.
3. **Prediction** – Use the modified model... to compute the probability of B, the consequence of the counterfactual.

Let us then examine one of Pearl's policy analysis examples that we can understand the definitions. The example consists of two equations:[26]

$$q \ \text{c=} \ b_1 p + d_1 i + u_1, \tag{3}$$

$$p \ \text{c=} \ b_2 q + d_2 w + u_2, \tag{4}$$

where q is the quantity demanded for some good, p is its price, i is income, w is the wage rate and u_1 and u_2 are error terms. Given the modularity assumption – each equation represents an autonomous mechanism – (3) and (4) constitute a causal model with Q and P as endogenous and U_1, U_2, I and W as the exogenous variables ($M =<\{U_1, U_2, I, W\}, \{Q, P\}, \{(3), (4)\}, P(u) >$).

Pearl considers the policy question, 'Given that the observed price is $P = p_0$, what would be the expected value of the demand Q had we controlled the price

[23] 'Acyclic' is a term from graph theory which basically means that there are no loops in the system.

[24] A direct cause is a cause that makes a contribution to the effect that is not mediated via other variables represented in the model.

[25] Theorem 7.1.7, p. 206

[26] pp. 215ff. The symbol 'c=' replaces Pearl's '=' and reads 'functionally equivalent and the variables on the right hand side cause the variables on the left hand side'.

to be $P = p_1$?'. The required probability, $P(Q_{P=p1} |\ P = p_0)$, can be evaluated as follows, using the three steps.

1. **Abduction** – Update $P(u_1)$ by the evidence $P = p_0$ as well as $I = i$, $W = w$ to obtain $P(u_1|P = p_0, I = i, W = w)$.

2. **Action** – Modify M by the action $do(P = p_1)$. That is, formulate the submodel $M_{P\ =p1}$:

$$q \leftleftharpoons b_1 p + d_1 i + u_1,$$
$$p = p_1.$$

3. **Prediction** – Use the modified model to compute the probability of Q. This yields for the expected value of Q:

$$E(Q_{P=p1}|p_0, i, w) = b_1 + p_1 + d_1 i + E(U_1|p_0, i, w).$$

One of the advantages of Pearl's semantics is that it ties in very nicely with purely philosophical accounts. In particular, it can be shown that under certain conditions Pearl's and Lewis's semantics yield *the same* results (for recursive systems).[27] In order to achieve this equivalence, the counterfactual $Y_x(u) = y$ Pearl's notation is equated with Lewis's $A \mathrel{\square\!\!\rightarrow} B$ (and A with the proposition that $X = x$ holds and B with the proposition that $Y = y$ holds).

In Lewis's account, the counterfactual antecedent is brought about 'by miracle'. That is, the laws that are responsible for A to obtain are broken and A is brought about *ex nihilo*. Equivalently, Pearl explicitly breaks all causal laws that have X as an effect and replaces these laws with the constant $X = x$.

What seems to be an advantage from a philosophical point of view is detrimental from the point of view of the policy maker. For Pearl's semantics to work, it is required that the law for each variable in the system (P and Q in the example) be separately manipulable, and he reads the counterfactuals as supposing that the implementations can be represented by changes in exactly the laws governing the antecedent and no others. To evaluate a counterfactual, we model the implementation in one special way: by destroying one set of causal laws, replacing it with another set and leaving the rest of the system intact.[28]

Do all socio-economic systems of interest for policy making behave like this? We believe not. We shall illustrate that by pointing to one large class of systems

[27] Pearl 2000, pp. 238ff.

[28] It is of course always possible to evaluate a counterfactual according to Pearl's method *assuming* the truth of a model with the right properties (as we have done in the example). Thus, whether or not real socio-economic systems have these properties does not matter. Pearl is explicit, however, that he understands causality and related notions to append to the world, rather than a model (or our language, say). Models represent real features of the world. For policy considerations, then, we want the model to be true of the world and therefore the system in the world, not just the model, must have the right properties.

relevant for policy making familiar in economics. These are systems where agents' expectations play a role in determining the relations between the control variable and the target variable (the well-known 'rational expectations' models represent a special case of these). Consider a simple model of the money market:

money demand:

$$m_t = p_t + y_t - \lambda i_t \tag{5}$$

money supply:

$$m_t = m^* + \gamma(y^* - y_{t-1}) + \upsilon_t, \tag{6}$$

where m = money, p = price level, y = real output, i = nominal interest rate, m^* = exogenous money supply, y^* = potential output and υ = white noise. This we present this model for expositional purposes only, we omit the goods side of the economy, which contains the expectations parameters.

Let us consider the counterfactual question 'What would the value of the price level p have been, had the money supply been m_c?' How would we model that counterfactual à la Pearl? He asks us first to replace the feedback mechanism (6) with a constant law:

$$m_t = m_c, \tag{7}$$

where the m_c defines some constant value of m. What would the agents in the economy make out of it? As far as we know, there is no uniform answer in the literature but here are a number of possible replies.

First and foremost, the reply will depend on what the agents know. It seems that the most natural reply to Pearl's question would be that despite the central bank's regime change, the agents would believe that (6) is still intact. Since m_c is a possible realisation of (6), *viz.* in case

$$\upsilon_t = m_c - m^* - \gamma(y^* - y_{t-1}),$$

the agents might still use (5)-(6) in their decision making. The point is that Pearl wants us to model the shift from the actual to the counterfactual situation minimally. Thus he replaces one law in the system with another one. If that means that the agents' expectations remain the same the central bank can change as much as it wants, agents' expectations would be formulated on the basis of (5)-(6) rather than (7). In this case, we would have to formulate Pearl's question more precisely as 'What would the value of the price level p have been, had money supply been m_c due to a realisation of the error term of $\upsilon_t = m_c - m^* - \gamma(y^* - y_{t-1})$?'. In order to evaluate this counterfactual, we would just update the agent's beliefs using (5)-(6) and the realisation of υ_t. Pearl's semantics would not apply.

Alternatively, we could mean the counterfactual to say that central bank policy has now changed (temporarily? permanently?) to a fixed policy regime. Fixed-policy regimes, however, are usually not modelled as (7) but as a mix of a

deterministic term plus a random component. Economic agents are usually not held to believe that the central bank can perfectly control the money supply. Thus we have, *e.g.*,

$$m_t = m_c + \varepsilon_t. \tag{8}$$

This comes closest to Pearl's suggestion. It would answer the question "What would the value of the price level p have been, had money supply been m_c due to (a) a credible switch of monetary policy to a fixed regime and (b) the realisation of $\varepsilon_t = 0$?"

Another alternative is to have the agents believe that the central bank has changed its policy to a so-called regime-switching system. This could be represented by

$$m_t = \begin{array}{l} m^* + \gamma(y^* - y_{t-1}) + \upsilon_t \text{ with probability } q \\ \\ m_c + \varepsilon_t \text{ with probability } (1-q). \end{array} \tag{9}$$

The counterfactual question we answer with this model would be very similar to the preceding one but we would have 'regime switching system' instead of 'fixed policy regime'. None of the preceding models models the expectations formation process of the agents very deeply. In particular, one can assume that the agents realise that the central bank's decision to set the value for the money supply is itself an outcome of a rational procedure. For example, they might believe that the central bank aims at minimising the following loss function:

$$L = E_{t-1}(p_t - z_{t-1})^2, \tag{10}$$

where z is an exogenous target value. The central bank would then set the money supply m_c such that $E_{t-1}(p_t) = z_{t-1}$. A change in m_c, then, would imply that the target variable z has changed.

The moral of this story is that if expectations matter, Pearl's semantics might not be applicable. Pearl assumes that we can change each law on its own. However, if some laws in a system contain expectations, the *actual* changes may be of little relevance. What matters more is what agents can be made to believe. And in order to make them believe the right thing, many more things might have to change than the one law Pearl envisages.

This is not, however, a problem confined to cases of rational expectations but a general one for Pearl's account. Pearl evaluates one particular kind of implementation-specific counterfactual, where the counterfactual antecedent is brought about by a precise incision that changes exactly the laws governing the counterfactual antecedent and nothing else *at all* (except what follows causally from just that difference). When we consider implementing a policy, we want to know what would happen were the policy really set in place – and this may well involve a variety of changes beyond those Pearl admits.

The problem is not with implementations that involve complex changes each of which can be represented in the overall scheme as a change in the value of one of the variables. Although we have not described it here, Pearl offers a detailed account of how to deal both with cases where our actions are complex – they involve changes in a number of different variables at once, and also where our complex actions may be conditional in specified ways on the values that other variables take. The problem is rather with implementations that result in changes in some of the causal laws that describe the system.

We might most easily locate this problem in Pearl's characterisation of a causal model. Recall, a causal model for Pearl contains a set of equations, one for each effect. The effect is to be written on the left-hand side; on the right is a full set of direct causes of that effect, where the values of these causal variables fix the value of the effect. Note first that this means that Pearl's scheme cannot be used to evaluate counterfactuals in systems where causes may act purely probabilistically. Nor can it deal with systems where the causal laws themselves may be affected by our policies, for instance, where the causal laws are tied together so that if one changes, so too will others, or so too probably will others.

This is a familiar phenomenon and we are not always out of our depths in dealing with it. Occasionally we understand what will happen to the laws of the system as we implement different kinds of change and can provide a scientific account of it. This is one of the central aims of rational expectations theory, which is why we have used a rational expectations case as an example. Rational expectations methods cannot be encompassed by Pearl's scheme. But this is only an example. The general problem with Pearl's scheme is that it provides no place to encode information that we might well have about how the changes we envisage will affect the causal laws by which the system operates. His characterisation of what constitutes a causal model is too narrow.

We should note that there is one quick solution to this problem that will not do: simply add more causal laws to express the information needed. In its first preliminary form the answer will not do because we need information about how changes in causal laws are correlated with each other and it does not make sense to write down causal laws that take causal laws themselves as both antecedents and consequents.

A different version of the answer supposes that where these problems arise it must always be because we are focusing on the wrong set of causal laws, a set of causal laws at the 'wrong level'. This view takes as a paradigm the situation in which there is some set of fundamental laws that have no connections with each other: each can change separately. From these, given certain specified boundary conditions, some further less fundamental laws can be derived. Under this picture, if there is a change in some targeted derivative law (keeping the boundary conditions fixed), there must necessarily be a change in the fundamental laws; this in turn might well lead to changes in other derivative laws beyond the one targeted. In this kind of situation, so the answer goes, counterfactuals should be evaluated relative to a causal model that describes the fundamental laws, not one that employs the derivative.

There are a number of problems with this way of defending the universal applicability of Pearl's scheme. The first concerns what the character of economic laws really is. We agree that the story of two tiers of causal laws, one more fundamental than the other, does show how correlations among laws can come about. But we have no reason to think that it is the only way. Moreover, whether there is one tier or many, it does not seem likely– no matter how far 'down' we go – that the principles for economic systems need ever have the kind of independence from each other that Pearl's scheme demands[29].

The second set of problems concern the more realistic issues of how best to deploy what we know. Even if the two-tier story is the right account for why changes in certain economic principles occur in tandem, it is very often much harder to learn about the 'more fundamental' tier given even our best methods for empirical inference in economics. Nor do we need to know the principles at this level to evaluate our counterfactuals. What we need to know and to encode in our causal models is how the principles we propose using to make counterfactual judgements are likely to change given our envisaged implementations, and that is something we often come to learn, or to have good bets about, without having to calculate it from a 'more fundamental' theory.

Last, we should notice that rational expectations theory itself does not fit Pearl's scheme. Notoriously the theory does offer a two-tier account of why certain 'observational' regularities may well not be stable as we try to use them to implement policy. The observational regularities are supposed to be a consequence of the behaviour of rational agents and those agents will – because they are rational – change their mode of behaviour if they foresee that a new policy will be implemented. Thus the observational regularities that arise under the older modes of behaviour may well no longer obtain were the policy to be implemented (or were agents to expect it to be implemented).

Nevertheless the methods suggested by rational expectations theory for calculating what would happen were a new policy undertaken are more complex than those proposed by Pearl. We do not just replace the law for the policy variable with a new one setting the value at the proposed level, then deduce from the 'fundamental' laws in their original unchanged form with parameters fixed what the values of the target variables will be. For there is an interaction envisaged between the two tiers – a 'self-consistency' requirement is invoked. The variables that appear in the 'fundamental' laws include expectations (in the sense of beliefs and predictions) that the agents have about macro variables; the values of these are supposed to match the expectations (in the statistical average sense) of those macro variables in the less fundamental laws. This provides a method – albeit not formalised – for calculating simple (non-complex, non-conditional) counterfactuals. But so far as we can see, it cannot be fitted into Pearl's scheme.

Why does Pearl focus on this one kind of implementation, where variables change as the counterfactual antecedent is implemented, but no laws, excepting those governing the counterfactual antecedent? One reason is to resolve the

[29] See *e.g.* Estrella and Fuhrer 1999.

problems of ambiguity. As we discussed in Section 3 ordinary language counterfactuals are open. They have different truth values depending on how they are supposed to be implemented. We have advocated disambiguating the counterfactual before trying to evaluate it; so too does Pearl. His detailed work on how to assess counterfactuals with complex and conditional antecedents is testimony to that.

Pearl is also worried about the problems of ambiguity that arise if laws can change. His solution here is to treat all counterfactuals the same. He insists that the laws of a causal model can be changed one at a time and that the right reading of the counterfactual is the one that supposes that that is just what happens: the laws governing the antecedent change and only those laws. This will work sometimes, but, as we have argued, it does not allow us to answer all the 'What if?' questions we want to ask. On the other hand, Pearl's is the most complete and well developed formal apparatus available for assessing counterfactuals in science. This is why we are keen to bring this issue to the fore.

One of the tasks that Pearl's particular semantics undertakes is that recommended by the Hume programme, that is, the elimination of 'problematic' causal notions in favour of non-causal ones. For example, Pearl defines *the causal effect* of one variable on another in terms of counterfactuals and then shows how to calculate the causal effect from the probabilities of propositions that employ neither causal nor counterfactual notions. As we noted, this is just the kind of programme that philosophers have been heavily engaged in; in particular Pearl's account is very similar to that of David Lewis[30]. Pearl's account like Lewis's is skewed towards the job of evaluating counterfactuals that can deliver the right verdicts on the targeted causal concepts. Correlatively it fails to provide the right answers to large numbers of genuine 'What if?' questions we want to ask.

What then of our claim above that Pearl provides a proof that his semantics is indeed a semantics for counterfactuals? The proof is in the form of a representation theorem. He first provides some independent characteristics that counterfactuals are supposed to satisfy, then shows that these characteristics hold of the propositions his semantics evaluates and, conversely, that for any set of propositions for which these characteristics hold, there is a causal model renders them true under his semantics.

The trouble is with Pearl's characteristics for counterfactuals. One is called *composition*. Pearl describes it this way: 'Composition states that, if we force a variable (W) to a value w that it would have had without our intervention, then the intervention will have no effect on other variables in the system.'[31] Composition is fine when antecedents are implemented by miracles of the kind employed by Pearl and Lewis. But it is not a characteristic of counterfactuals in general. Most ways of implementing a counterfactual antecedent will produce knock-on effects different from what would have obtained had the antecedent

[30] See Lewis 1973, 1979 and 1986.

[31] Pearl 2000, p. 229

occurred without the implementation. That is particularly true of the kinds of real life implementations available to us in complex policy situations.

Pearl's scheme can serve as a model for what is needed. It has the three requisite ingredients for a semantics for counterfactuals and it is both rich in detail and formally set-up to allow for the proof of a variety of important and interesting results. But it itself does not do the job of providing a semantics for genuine 'What if?' counterfactuals.

5 Input Counterfactuals and Output Counterfactuals: James Heckman

Counterfactuals are an important topic in economic methodology today, especially since they have been championed by Nobel prize winner James Heckman[32]. Superficially the way the counterfactuals are pictured to work by Heckman is much as we have urged. A causal model is postulated and rules are supplied for how to calculate answers to specific 'What if?' questions from the model. But when it comes to setting counterfactuals to use there is a big difference. For the counterfactual analysis provided by Heckman cannot in general answer 'What if?' questions of the kind we pose in planning, policy and evaluation. Nor does Heckman claim that it can. The analysis is offered, rather, as in the philosophy literature, as a causal surrogate. We wish to underline the fact that the two jobs are different. We cannot assume that what serves for the one job will serve for the other.

This does not mean, however, that the causal-surrogate counterfactuals are irrelevant for policy and planning. We shall point out that they can serve as *inputs* for constructing the causal models employed in generating the genuine policy counterfactuals needed for planning as outputs. This will require, however, a weaker interpretation than that put on the causal-surrogate counterfactuals.

As in the philosophy literature, Heckman's counterfactuals are offered as a way of defining causal concepts in terms of presumably less problematic non-causal concepts. We propose to view them instead as a way for finding out about independently understood causal relations in special situations. Entirely separate sets of considerations may then tell us whether we can export knowledge about these causal relations to new situations, for which we need to construct causal models and generate counterfactual outputs.

What kinds of considerations are required? The distinction between internal and external validity is of use here. A causal claim is internally valid if it is correct for the experimental system in which it was established. By contrast, the claim is externally valid if in addition it is correct for (possibly non-experimental) situations outside the original system. There is no systematic answer available to the question of how to establish external validity. We make only one small contribution here. The counterfactuals that are on offer as causal surrogates have little external validity when they are conceived as coun-

[32] Heckman 2000 and 2001

terfactuals. Indeed they very often do not even make sense in the new situation. They only become relevant if they can be taken as measures of causal relations that hold in the situations in which the causal-surrogate counterfactuals are evaluated and that might (or might not) hold in the target situation.

We have talked about the importance of counterfactuals to questions of policy. Similarly when we want evaluate the effectiveness of a trial programme we need answers to counterfactual questions: What if the programme had not existed? Or had existed in some other form? Or were set up more widely without the trial controls? These are just the kinds of questions Heckman considers in his applied work on the evaluation of labour market programmes, where he is at pains to point out that the question itself must be carefully formulated.

We have stressed the ambiguity in ordinary counterfactuals about how the antecedent will be implemented. Heckman points out other sources of ambiguity. We may for instance want to know what the wages of workers in the population at large would have been had the programme not existed; more commonly we end up asking what the wages of workers *in the programme* would have been. Or we may want to know what the GDP would have been without the programme. We also need to take care about the contrast class: do we want to know the difference between the results of the programme and those that would have occurred had no alternatives been present or the difference compared to other programmes, real or envisaged?

Heckman begins his treatment with *causal functions*. As with LeRoy's starting point, causal functions are a special kind of sparse causal model. The models describe special kinds of systems, systems that mimic experiments: 'Causal functions are ... derived from conceptual experiments where exogenously specified generating variables are varied... The specification of these hypothetical variations is a crucial part of model specification and lies at the heart of any rigorous definition of causality'.[33]

Heckman tells us three things about causal functions: i) They 'describe how each possible vector of generating variables is mapped into a resulting outcome', where the generating variables 'completely determine' the outcome.[34] ii) They 'derive from' – or better, we think, 'describe' – conceptual experiments. iii) Touching on questions of realism and of model choice, models involving causal functions are always underdetermined by evidence; hence, as Heckman sees it, causality is just 'in the head' since the models relative to which it is defined are just in the head. From this we take it that causal functions represent (a probably proper subset of) the causal principles under which these special experiment-like systems operate, where the right-hand-side variables – the ones Heckman calls the 'generating variables' – form a minimal complete set of causes of the quantity represented on the left[35] and where each cause can each vary independently of the others; *i.e.* they are variation-free.[36] This is why we say that Heckman's

[33] Heckman 2001, p. 14

[34] *ibid.*, p. 12

[35] Or, keeping in mind Heckman's view that causality is only relative to a model, the right-hand-side variables record what the model designates as causes.

[36] Formally, a set of variables $(x_1, x_2, \ldots, x_n)$ is variation-free iff $(x_1, x_2, \ldots, x_n) \in \mathcal{X}_1 \times \mathcal{X}_2$

starting point is a causal model.

Imagine that the causal function for an outcome y is given by

$$y \Leftarrow g(x_1, \ldots, x_n). \tag{11}$$

The *causal* or *counterfactual effect* (Heckman seems to use the terms 'causal effect' and 'counterfactual effect' interchangeably) of x_j on y fixing the remaining factors in the causal function is defined thus:

Causal effect of x_j *on* y:

$$[\Delta y/\Delta x_j = x'_j - x''_j] =_{df} g(x_1, \ldots, x'_j, \ldots x_n) - g(x_1, \ldots, x''_j, \ldots x_n). \tag{12}$$

Read in terms of counterfactuals we have here the evaluation, relative to a specific assignment of values to all $x_j, j \neq i$, of the difference between the value y would have were $x_i = x'_i$ and the value y would have were $x_i = x''_i$.

As Heckman insists, in order for this definition 'to be meaningful requires that the x_j can be independently varied when the other variables are fixed so that there are no functional restrictions connecting the arguments... it is thus required that these variables be variation-free'.[37] We shall call the counterfactual effect as thus defined a *Galilean effect* since it is just the kind effect we look for in a Galilean experiment, where a single cause is varied controlling for all other relevant features in order to observe the effect of that cause 'acting alone'.

Heckman considers simultaneous supply and demand equations as an example. For simplicity we can look at the specific equations that we have already considered in discussing LeRoy, where we have added the additional equilibrium constraint on price:

$$\begin{aligned} q_s &= \alpha_s + \alpha_{sp} p_s + \alpha_{sw} w \\ q_d &= \alpha_d + \alpha_{dp} p_d + \alpha_{di} i \\ q_s &= q_d = q \\ p_s &= p_d = p. \end{aligned} \tag{13}$$

Heckman points out that these equations do not fit Pearl's scheme since they are not recursive and hence Pearl's method for assessing counterfactuals will not apply. This fits with familiar remarks about these kinds of systems: p and q are determined jointly by exogenous factors. It seems then that it makes no sense to ask about how much a change in $p$will affect a change in q. To the contrary, Heckman points out. We can still assess causal efficacy using his definition – so long as certain 'exclusion' conditions are met.

Say we want to assess the causal/counterfactual effect of demand price on quantity demanded. We first look to the reduced form equations:

... $\times \mathcal{X}_n$.

[37] Heckman 2001, p. 18

$$q = (z_d, z_s) \tag{14}$$
$$p = (z_d, z_s),$$

where z_d is the vector of exogenous variables in the demand equations and z_s, those in the supply equations. In LeRoy's equations (13), $z_d = i$ and $z_s = w$. Heckman takes these to be causal functions, otherwise the causal model has not properly specified the 'exogenous' variables. That means that the exogenous variables are 'generating variables' for p and q and that they are variation free. Now the task is easy: "Assuming that some components of $[z_d]$ do not appear in $[z_s]$, that some components of $[z_s]$ do not appear in $[z_d]$, and that those components have a non-zero impact on price, one can use the variation in the excluded variables to vary [p_d or p_s in the reduced form equations] while holding the other arguments of those equations fixed".[38] The result (using the equality of p_d and p_s and of q_d and q_s) is

$$\partial q_d/\partial p_d - (\partial q/\partial z_s(e))/(\partial p/\partial z_s(e)), \tag{15}$$

where $z_s(e)$ is a variable in z_s that is excluded from z_d and that, as he puts it, 'has an impact on' p_d. In (13) this job can be done by w; the causal effect thus calculated of p_d on q_d is α_{dp}.

Notice how much causality is involved here. By definition we are supposed to be evaluating the change in q_d holding fixed all the factors in a causal function for q_d except p_d. What we actually do is hold fixed z_d while z_s varies. Presumably this is okay because z_s is a cause of p_d that can produce variations in p_d while $z_d$is fixed; and z_d being fixed matters because z_d constitutes, along with p_d, a minimal full set of causes of q_d. So when the exclusion condition is satisfied, the demand equation is a causal function and the counterfactual definition of causal effect is meaningful.

Now consider a slightly altered set of equations:

$$\begin{aligned} q_s &= \alpha_s + \alpha_{sp}p_s + \alpha_{sw}w + \alpha_{si}i \\ q_d &= \alpha_d + \alpha_{dp}p_d + \alpha_{di}i + \alpha_{dw}w \\ q_s &= q_d = q \\ p_s &= p_d = p. \end{aligned} \tag{16}$$

In this model the demand equation cannot be treated as a causal function and the question of the causal effect of demand price on quantity demanded is meaningless. This is true despite the fact that α_{dp} still appears in the equation and it still represents something – something much the same one would suppose – about the bearing of p_d on q_d. The intermediate case seems even stranger. Imagine that $\alpha_{sw} = 0$. Now α_{sp} measures a Galilean effect but α_{dp} does not.

[38] Heckman 2001, p. 36

We propose an alternative interpretation of what is going on. We begin with causal principles. One can be a realist about the principles of a causal model. They are correct if and only if they approximate well enough to the causal laws that govern the operation of the system in question. Heckman, it seems, is not a realist. But that does not matter here since he himself has introduced the notion of a causal function. A causal principle is just like a causal function but without the restriction that the causes (or 'generating variables') are variation free. We shall continue to restrict attention to linear causal models of the kind both LeRoy and Heckman use for illustration. Then define for any linear causal model, *the contribution a cause x_c makes to an effect x_e* $=_{df}$ the coefficient of x_c in any causal principle for x_e in the model.[39] It is trivial to show for any linear causal model that where Heckman's measure for the causal/counterfactual effect of x_c on x_e applies, it has the same value as the contribution x_c makes to x_e.

Given this characterisation the contribution of p_d to q_d is the same in (13) and (16). What is different is that in (13) we have a particular way to find out about it that is not available in (16). (13) is what we call *an epistemically convenient system*[40]: having an epistemically convenient system implies among other things that it is possible to find out what a cause, x_c, contributes to an effect, x_e, in one particular simple way – hold fixed all the other contributions that add up to make the effect the size it is; then vary the cause and see how much x_e varies. Any difference has to be exactly the contribution that x_c adds. This does not mean, however, that for systems where this independent variation is not possible, all is lost. There are hosts of other legitimate ways of defending claims about the size of causal contributions that apply both in systems with independent variation and in ones without.

There are two advantages to the account that takes Heckman's causal surrogate counterfactuals as measures of causal contributions rather than as mere counterfactuals.

First, few systems we confront are governed by principles in which the causes are variation free. The vast majority are not. In these systems Heckman's counterfactuals are irrelevant.

Second, even if we are studying a system where the causes are variation free, there is a puzzle about why we should wish to ask just these implementation-specific questions. If we are thinking of setting policy or evaluating the success of some programme in the system, then these, with their special method of implementation, might be relevant sometimes. But there is no necessity to implement policies in the single way highlighted by Heckman; generally we would want to consider a variety of different methods of implementation and frequently to assess implementation-neutral counterfactuals as well. Even where causes are variation free, the counterfactual changes that Heckman studies generally have no privileged role.

[39] This supposes that all principles in the model with x_c on the right-hand-side and x_e on the left will have the same coefficient. This will be the case given a proper statement of 'transitivity' and the definitions for the form of causal principles sketched in Cartwright forthcoming.

[40] For a definition see Cartwright forthcoming.

There are two familiar enterprises where they do have a special role. The first is in trying to determine if, and to what degree, one factor contributes causally to another. In an epistemically convenient system we can ask Galilean-type counterfactual questions; and the answers we obtain will double as answers to our causal questions. They are a tool for finding out answers to our causal questions. But note that they are only a tool for finding out about causes in special epistemically convenient systems. For other systems we cannot even ask these counterfactual questions, let alone let the answers to them supply our causal answers as well.

The other is in Heckman's own field, evaluation. In setting up new programmes, we might try to set them up in such a way that the causal contribution they make to the result can be readily disentangled from the contribution of other factors. Of particular concern are other factors that might both contribute to the effect independently of the programme and also make it more likely that an individual entered (or failed to enter) the programme. If we can arrange the setup of our programme so that it is epistemically convenient, then again we can ask Galilean counterfactual questions – 'What difference would there be in outcome with the programme present versus the programme absent, holding fixed all other contributions to the outcome?' And again these counterfactual questions will tell us the contribution the programme makes, since in these circumstances the difference in outcome between when the programme is present and when it is absent must be exactly the contribution the programme makes. So we can use information about Galilean effects to learn about the causal contributions of the programme we set up. Still, all we learn is about that programme in those special epistemically convenient circumstances.

In either case, whether it be experimental systems or programme set-ups that we engineer to make the measurement of causal contributions easy, we need to ask, why should we be interested in causal contributions in these special – and rare – kinds of systems? The answer is clear. Generally we want this information because we hope it will tell us something about causal contributions in other systems. But we confront here the familiar problem of internal and external validity. In an epistemically convenient (linear) system, using counterfactual differences as a measure of causal contributions is provably valid: internal to the situation this method is bound to give us correct results about the question of interest. But nothing said in this discussion bears on external validity: when will the results that we can be sure are correct in an epistemically convenient system hold elsewhere?

So how do these odd Galilean counterfactuals bear on more useful 'What if?' counterfactuals? In a very indirect way, it seems. We can use Galilean counterfactuals to tell us about causal contributions in Galilean experiments. Then, to the extent that we can expect the facts about causal contributions to remain stable across situations of interest, we can use the information about causal contributions in Galilean experiments to help build causal models for new non-Galilean situations. And we can use the causal models so constructed for these new situations to answer real 'What if?' questions that we want to ask about these new situations. Galilean counterfactuals are one of many inputs in

a multi step process that yields as outputs counterfactuals of immediate use in policy and planning.

It is important to stress, however, that assumptions about when information about causal contributions learned in one setting will obtain in others are not justified by anything we have discussed so far, and in particular not by any information about counterfactuals of the kinds we have explored. Showing that results on causal contributions have external validity – and how far and of what kind – requires a different methodology altogether. This is one of the reason we have been at pains to distinguish input counterfactuals from output counterfactuals.

Recent work in economics like that of LeRoy and Heckman provides some clear formal characterisations of how to evaluate special counterfactuals with very particular assumptions about their methods of implementation. But they leave us a long way from any characterisation of counterfactuals whose implementations are the ones we are generally interested in. For we have no good rules for how to judge external validity; that is, for when we can export what we learn from a Galilean effect to build a causal model for a non-Galilean situation. And then, as we saw in discussing Pearl, we have only a thin start on how to use a causal model to answer the 'What if?' questions we ask for planning and policy.

6 Conclusion: Disambiguate Before You Evaluate

The lesson we want to draw from the preceding discussion can be captured by the slogan 'Disambiguate before you evaluate'. All three accounts we have been looking at provide unambiguous semantics for counterfactuals. But they do so at the cost of seriously constraining the range of admissible counterfactual questions. More importantly, they often provide semantics for counterfactuals that have no relevance for policy analysis whatsoever. What we urge instead is to disambiguate the question first and subsequently use semantics that are tied closely to precisely that kind of question.

First and foremost, we will need to know whether the counterfactual at stake is Heckman-style Galilean counterfactual of the kind we have discussed in Section 5 or a genuine policy counterfactual as discussed in Sections 2-4. A Galilean counterfactual helps us to evaluate the contribution of a causal factor to a quantity of interest in very specific – that is, epistemically convenient – systems. Without further tests, the results of this evaluation tell us nothing beyond the experimental system at hand. But we may be lucky and find that test results are exportable or 'externally valid'. In this case, we can use the knowledge gained in one particular system as a building block for the evaluation of a genuine policy counterfactual.

Let us assume that we have knowledge of all causal laws governing a particular socio-economic system. If we now want to evaluate a genuine policy coun-

terfactual, we have to decide whether it is meant in an implementation-specific or neutral sense. Implementation-neutral counterfactuals are true no matter how the antecedent is brought about. These come in particularly handy in case we don't know how to implement the antecedent. But this kind of epistemic uncertainty is neither necessary nor sufficient for evaluating counterfactuals in an implementation-neutral way. On the one hand, we might just want to know what the answer to our counterfactual question was in case it didn't matter how we implemented the antecedent. On the other hand, there are other strategies available for situations of epistemic uncertainty. And these are important if the causal structure of the socio-economic system that we contemplate does not lend itself to implementation-neutral evaluation (Recall for instance that there was no answer to LeRoy's question for system (1), where the price variable appeared in both equations).

Turning to implementation-specific counterfactuals, finally, once more both metaphysics and epistemology matter. For systems which are characterised by the axioms of Pearl's formalism, we can use his semantics and implement the antecedent without disturbing any other relationship in the system. But for many socio-economic systems that will not do, so we may have to use the more involved (and less formal) modelling strategies of, say, rational expectations economics. In both cases we assume that we have a clear idea of how the antecedent is to be brought about.

If not, we must fall back on alternative strategies to deal with epistemic uncertainty. Though we may not know for sure where the money for the new schooling programme comes from, we may know that it is far more likely to be raised by a tax increase than by cutting defence. Thus we can use our best guess as to the probability of each possible implementation and weigh the results accordingly.

The main message of our musings about the use counterfactuals in policy advice is, then, to make the underlying assumptions as transparent as possible. Answering a counterfactual question can amount to a myriad of things. We believe we had better be clear about precisely what issues are involved and give answers that are tailored to the specific case at hand, rather than provide a semantics that yields results across the board but in so doing fails to resolve the question we are in fact interested in.

References

Cartwright, Nancy (1983), *How the Laws of Physics Lie*, Oxford: OUP.

Cartwright, Nancy (1989), *Nature's Capacities and Their Measurement*, Oxford: OUP.

Cartwright, Nancy (forthcoming), "Two Theorems on Invariance and Intervention", *Philosophy of Science*.

Edgington, Dorothy (1995), "On Conditionals", *Mind* **104**, 235-329.

Estrella, Arturo and Jeffrey Fuhrer (1999), "Are 'Deep' Parameters Stable? The Lucas Critique as an Empirical Hypothesis", *Federal Reserve Bank of Boston Working Paper* 99-4.

Hamilton, James (1994), "Measuring the Liquidity Effect", *American Economic Review* **87**:1, 80-97.

Hausman, Daniel (1998), *Causal Asymmetries*, Cambridge: CUP.

Heckman, James (2000), "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective", *Quarterly Journal of Economics* **115**:1, 45-97.

Heckman, James (2001), "Econometrics Counterfactuals and Causal Models", Keynote Address *International Statistical Institute*, Seoul, South Korea.

Holland, Paul (1988), "Statistics and Causal Inference", *Journal of the American Statistical Association* **81**:396, 945-960.

Kyburg, Henry and Mariam Thalos (2003), *Probability Is the Very Guide of Life*, Chicago and La Salle (Ill.): Open Court.

LeRoy, Stephen (2003), "Causality in Economics", MS, University of California, Santa Barbara.

Lewis, David (1973), "Causation", *Journal of Philosophy* **70**, 223-67. Reprinted in Sosa, Ernest and Michael Tooley, *Causation*, Oxford: OUP (1993), 193-204.

Lewis, David (1979), "Counterfactual Dependence and Time's Arrow", *Noûs* **13**:4, 455-76.

Lewis, David (1986), "Postscripts to 'Causation"', *Philosophical Papers Vol. 2*, Oxford: OUP, 172-213.

Lewis, David (2000), "Causation as Influence", *Journal of Philosophy* **97,** 182-197.

Pearl, Judea (2000), *Causation: Models, Reasoning and Inference*, Cambridge: CUP.

Reiss, Julian (forthcoming), "Review of Kevin Hoover's *Causality in Macroeconomics*", *Forum for Social Economics.*

Rubin, Donald (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology* **66**, 688-701.

Simon, Herbert (1953), *Causal ordering and identifiability.* In Hood and Koopmans (eds), *Studies in Econometric Method.* Cowles Foundation monograph No. 14.

Simon, Herbert (1954), "Spurious Correlation: A Causal Interpretation", *Journal of the American Statistical Association* **49**, 469-492.

Spirtes, Peter, Clark Glymour and Richard Scheines (2001), *Causation, Prediction and Search*, Cambridge (Mass.): MIT Press.

Stalnaker, Robert (1968), "A Theory of Conditionals", in *Studies in Logical Theory*, *American Philosophical Quarterly* Monograph Series, 2, Oxford: Blackwell, 98-112.